VŠB - Technická univerzita Ostrava
Fakulta metalurgie a materiálového inženýrství
Katedra ekonomiky a managementu v metalurgii

Zadání bakalářské práce

Student: **Jakub Velký**

Studijní program: B3922 Ekonomika a řízení průmyslových systémů

Studijní obor: 6208R123 Ekonomika a management v průmyslu

Téma: Data mining a jeho možnosti využití v podniku
Data Mining and Possibilites of it's Aplication in Company

Zásady pro vypracování:

Práce se zabývá obecně data miningem. Na základě teoretických poznatků uveďte a porovnejte jednotlivé způsoby a možnosti využití data miningu. Navrhněte konkrétní oblasti uplatnění a způsobu použití data miningu v podniku.

Seznam doporučené odborné literatury:

Rud, P. O. Data mining. Praha: Computer Press, 2001
Berka, P. Dobývání znalostí z databází. Praha: Academia, 2003.
Lacko, L. Databáze: datové sklady, OLAP a dolování dat. Praha: Computer Press, 2003.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Martin Lampa, Ph.D.**

Datum zadání: 30.11.2009
Datum odevzdání: 30.04.2010

VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA Fakulta metalurgie a mat. inženýrství 1 OSTRAVA

prof. Ing. Ivo Janík, CSc.
vedoucí katedry

prof. Ing. Ľudovít Dobrovský, CSc., Dr.h.c.
děkan fakulty

Zásady pro vypracování bakalářské práce

I.

Bakalářskou prací (dále jen BP) se ověřují vědomosti a dovednosti, které student získal během studia, a jeho schopnosti využívat je při řešení teoretických i praktických problémů.

II.

Uspořádání bakalářské práce:

1. Titulní list + zásady pro vypracování BP
2. Prohlášení + místopřísežné prohlášení
3. Abstrakt + klíčová slova česky a anglicky
4. Obsah BP
5. Textová část BP
6. Seznam použité literatury
7. Přílohy

ad 1) Titulním listem je originál zadání BP, který student obdrží na své oborové katedře. Za titulním listem následují tyto „Zásady pro vypracování bakalářské práce".

ad 2) Prohlášení + místopřísežné prohlášení napsané na zvláštním listě (student jej obdrží na své oborové katedře) a vlastnoručně podepsané studentem s uvedením data odevzdání BP. *V případě, že BP vychází ze spolupráce s jinými právnickými a fyzickými osobami a obsahuje citlivé údaje, je na zvláštním listě vloženo prohlášení spolupracující právnické nebo fyzické osoby o souhlasu se zveřejněním BP.*

ad 3) Abstrakt a klíčová slova jsou uvedena na zvláštním listě česky a anglicky v rozsahu max. 1 strany pro obě jazykové verze.

ad 4) Obsah BP se uvádí na zvláštním listě. Zahrnuje názvy všech očíslovaných kapitol, podkapitol a statí textové části BP, odkaz na seznam příloh a seznam použité literatury, s uvedením příslušné stránky. Předpokládá se desetinné číslování.

ad 5)

Textová část BP obvykle zahrnuje:

- Úvod, obsahující charakteristiku řešeného problému a cíle jeho řešení v souladu se zadáním BP;
- Vlastní rozpracování BP (včetně obrázků, tabulek, výpočtů) s dílčími závěry, vhodně členěné do kapitol a podkapitol podle povahy problému;
- Závěr, obsahující celkové hodnocení výsledků BP z hlediska stanoveného zadání.

BP nemusí obsahovat experimentální (aplikační) část.

BP bude zpracována v rozsahu min. 25 stran (včetně obsahu a seznamu použité literatury).

Text musí být napsán vhodným textovým editorem počítače po jedné straně bílého neleského papíru formátu A4 při respektování následující **doporučené** úpravy - písmo Times New Roman (nebo podobné) 12b; řádkování 1,5; okraje – horní, dolní – 2,5 cm, levý – 3 cm, pravý 2 cm. Fotografie, schémata, obrázky, tabulky musí být očíslovány a musí na ně být v textu poukázáno. Budou zařazeny průběžně v textu, pouze je-li to nezbytně nutné, jako přílohy (viz ad 7).

Odborná terminologie práce musí odpovídat platným normám. Všechny výpočty musí být přehledně uspořádány tak, aby každý odborník byl schopen přezkoušet jejich správnost. U

vzorců, údajů a hodnot převzatých z odborné literatury nebo z praxe musí být uveden jejich pramen - u literatury citován číselným odkazem (v hranatých závorkách) na seznam použité literatury.
Nedostatky ve způsobu vyjadřování, nedostatky gramatické, neopravené chyby v textu mohou snížit klasifikaci práce.

ad 6) BP bude obsahovat alespoň 10 literárních odkazů, z toho nejméně 3 v některém ze světových jazyků.
Seznam použité literatury se píše na zvláštním listě. **Citaci literatury je nutno uvádět důsledně v souladu s ČSN ISO 690.** Na práce uvedené v seznamu použité literatury musí být uveden odkaz v textu BP.

ad 7) Přílohy budou obsahovat jen ty části (speciální výpočty, zdrojové texty programů aj.), které nelze vhodně včlenit do vlastní textové části, např. z důvodu ztráty srozumitelnosti.

III.

Bakalářskou práci student odevzdá ve dvou knihařsky svázaných vyhotoveních, pokud katedra garantující studijní obor neurčí jiný počet. Vnější desky budou označeny takto:

nahoře: *Vysoká škola báňská - Technická univerzita Ostrava*
Fakulta metalurgie a materiálového inženýrství
Katedra . . . . . .

uprostřed: *BAKALÁŘSKÁ PRÁCE*

dole: *Rok* *Jméno a příjmení*

Kromě těchto dvou knihařsky svázaných výtisků odevzdá student kompletní práci také v elektronické formě do IS EDISON včetně abstraktu a klíčových slov v češtině a angličtině.

IV.

Bakalářská práce, která neodpovídá těmto zásadám, nemůže být přijata k obhajobě. Tyto zásady jsou závazné pro studenty všech studijních programů a forem bakalářského studia fakulty metalurgie a materiálového inženýrství Vysoké školy báňské – Technické univerzity Ostrava od akademického roku 2009/2010.

Ostrava 30. 11. 2009

Prof. Ing. Ľudovít Dobrovský, CSc., Dr.h.c.
děkan fakulty metalurgie a materiálového inženýrství
VŠB-TU Ostrava

PROHLÁŠENÍ

Prohlašuji, že

- jsem byl seznámen s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. - autorský zákon, zejména §35 - užití díla v rámci občanských a náboženských obřadů, v rámci školních představení a užití díla školního a §60 - školní dílo.

- beru na vědomí, že Vysoká škola báňská - Technická univerzita Ostrava (dále jen VŠB - TUO) má právo nevýdělečně ke své vnitřní potřebě bakalářskou práci užít (§35 odst. 3).

- souhlasím s tím, že jeden výtisk bakalářské práce bude uložen v Ústřední knihovně VŠB - TUO k prezenčnímu nahlédnutí a jeden výtisk bude uložen u vedoucího bakalářské práce. Souhlasím s tím, že údaje o bakalářské práci budou zveřejněny v informačním systému VŠB-TUO.

- bylo sjednáno, že s VŠB - TUO, v případě zájmu z její strany, uzavřu licenční smlouvu s oprávněním užít dílo v rozsahu §12 odst. 4 autorského zákona.

- bylo sjednáno, že užít své dílo - bakalářskou práci nebo poskytnout licenci k jejímu využití mohu jen se souhlasem VŠB - TUO, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly VŠB - TUO na vytvoření díla vynaloženy (až do jejich skutečné výše).

- beru na vědomí, že odevzdáním své bakalářské práce souhlasím s jejím zveřejněním podle zákona č. 111/1998Sb., o vysokých školách a o změně a doplnění dalších zákonů (Zákon o vysokých školách) bez ohledu na výsledek její obhajoby.

- Místopřísežně prohlašuji, že jsem celou bakalářskou práci vypracoval samostatně.

V Ostravě 28.4.2010

JAKUB VELIKÝ
jméno a příjmení studenta

OBĚŽNÁ 1, KOZMICE 74711
adresa trvalého pobytu studenta

Abstrakt: Cílem této práce by mělo být seznámení se s pojmem data mining, jeho další členění (neuronové sítě, rozhodovací stromy) a možnosti využití např. pro podnik, nebo ve spojení s CRM (řízení vztahů se zákazníky) je to například analýza prodeje, nebo cílený marketing. Dále je zde uvedena metodologie CRISP-DM, která dokáže řešit data miningové projekty rychleji a efektivněji a méně nákladně. Další část této práce se zaobírá daty a to typy dat, zdroji dat a klasifikaci dat.

Abstract: The aim of this work should be familiar with the concept of data mining and its subdivisions (neural networks, decision trees), and the possibilities of such an undertaking, or in conjunction with the CRM (customer relationship management) is an example of sales analysis, or targeted marketing. Furthermore, it is listed as CRISP-DM methodology that can address the data mining projects faster and more efficiently and less expensively. Another part of this thesis deals with data and data types, data sources or data classification.

Klíčová slova: data mining, datový sklad, databáze, data

Keywords: data mining, Data Warehousing, database, data

PODĚKOVÁNÍ

Děkuji Ing. Martinovi Lampovi Ph.D. za hodnotné rady a odborné vedení během mé práce.

Obsah

1 Úvod 1

2 Data 3

2.1 Typy dat 3

2.1.1 Demografická data 3

2.1.2 Behaviorální data 3

2.1.3 Psychografická data 3

2.2 Zdroje dat 4

2.2.1 Interní zdroje 4

2.2.2 Externí zdroje 5

2.3 Klasifikace dat 5

2.3.1 Kvalitativní data 5

2.3.2 Kvantitativní data 6

2.3.2.1 Nominální data 6

2.3.2.2 Ordinální data 6

2.3.2.3 Intervalová data 6

2.3.2.4 Spojitá data 7

2.4 Výběr dat pro modelování 7

2.4.1 Data pro získání zákazníků 7

3 Dolování dat 8

3.1 Datová kvalita 9

3.2 Datový sklad 10

3.2.1 Podnikový sklad 11

3.2.2 Datové tržiště 11

3.2.3 Virtuální sklad 11

3.3 Business intelligence 12

3.4 Analýza OLAP 13

3.4.1 Relační databázový model 13

3.4.2 Multidimenzionální databázový model 13

3.4.2.1 Multidimenzionální OLAP (MOLAP) 13

3.4.2.2 Relační databázový OLAP (ROLAP) 14

3.4.2.3 Hybridní OLAP (HOLAP) 14

3.5 Databáze 14

3.6 Relační databáze 14

3.7 Zneužití dat 15

3.8 Historie data miningu 15

4 Metody dolování dat 16

4.1 Neuronové sítě 16

4.2 Rozhodovací stromy 17

4.3 Logistická regrese 18

4.4 Kohonenovy mapy 19

5 Metodologie CRISP-DM 21

5.1 Porozumění problematice (definování cílů) 22

5.2 Porozumění datům 23

5.3 Příprava dat 24

5.4 Modelování 25

5.5 Vyhodnocení výsledků 25

5.6 Využití výsledků....26

6. Data mining v praxi 27

6.1 Data mining a CRM 27

6.2 Praktické rozdíly mezi vyhledáváním v databázích a data miningem 28

6.2.1 Data-mining a modelování 28

6.2.2 Vyhledávání v databázích 29

7. Závěr 31

8. Použitá literatura 32

9. Seznam obrázků 33

1 Úvod

V dnešní době je charakteristická exploze objemu dat sbíraných a ukládaných do databází. Disky pojmou stále větší množství dat, a proto neustále roste i objem ukládaných dat, ať už jsou pro nás užitečné nebo zbytečné. Dále roste také objem obchodních a průmyslových databází. V těchto datech je ukryto mnohem více informací, než lze z dat jednoduše vyčíst.

Služby (objednávky zásilkových služeb nebo cestovních kanceláří, popř. reservace jízdenek nebo letenek).

Bankovnictví (bankovní transakce, žádosti o úvěr, apod.).

Telekomunikace (informace o telefonním provozu a platbách za tento provoz, v případě mobilních telefonů obsahuje záznam informace také o poloze).

Státní správa (daňová přiznání, žádosti o sociální podporu, geografické informační systémy).

Podnik (odhalení potencionálního přechodu zákazníka ke konkurenci)

Zdravotnictví (zdravotní záznamy, informace pro zdravotní pojišťovny).

Zpracování dat z databází a datových skladů má v dnešním světě IT nejrůznější formy.

Tradiční přístupy analyzující data jsou dnes založeny na dotazovacích SQL, případně na technikách označovaných jako OLAP (On-line Analytical Processing), které často využívají uložení dat multidimensionálních databázích k rychlé prezentaci dat ve formě tabulek jako například (rok, čtvrtletí, měsíc versus kraje, okresy, obce). Tyto techniky umožňují udržovat přehled o okamžité pozici podniku, nebo rychlou přípravu finančních reportů. To vše lze realizovat také v rozsáhlých organizacích a během doby, která se před několika lety zdála být nesplnitelná.

Přesto je ještě mnoho úloh, na které tyto přístupy nestačí. Ve většině úloh není specifikován konkrétní dotaz na obsah databáze. Naopak cílem je, které údaje, nebo jejich kombinace jsou při komerčním využití dat z databáze důležité.

Pokud je obsahem tradičního databázového dotazu otázka, ve kterém kraji byl v tomto čtvrtletí nejúspěšnější prodejce zboží např. typu X? Pak dolování dat se snaží nalézt řešení problému: Jaké podmínky (skladba produktů, demografie cílové skupiny) zabezpečují dlouhodobě nejlepší výsledky prodeje.

Cílem této práce by mělo být seznámení se s pojmem data mining, jeho další členění a možnosti využití např. pro podnik.

Mezi největší výhody data miningu patří zvýšení efektivity práce a odhalování chyb. Díky tomu lze: Efektivně zvážit rizika, vyvíjet služby přímo na míru zákazníkům, optimalizovat produkty.

Zpracování dat z rozsáhlých databází a datových skladů má v dnešním světě IT velký význam. Data mining znamená mnoho různých postupů a algoritmů, které umožní odhalit a plně využít vztahy ukryté v datech.

2 Data

2.1 Typy dat

Data spadají do tří základních typů: demografický, behaviorální a psychografický. Každý typ má své výhody a nevýhody.

2.1.1 Demografická data

Obecně popisují charakteristiky osob či domácností. Mezi tento typ dat patří pohlaví, věk rodinný stav, příjem, vlastnictví domu, typ obydlí, úroveň vzdělávání, národnost a počet dětí. Demografická data mají řadu silných stránek. Jsou velmi stabilní, což je činí výborně použitelnými v prediktivních modelech. Charakteristiky jako rodinný stav, vlastnictví domu, úroveň vzdělání a typ obydlí se nemění tak často jako behaviorální data, jakými jsou zůstatky na účtech nebo názorově (psychografické) charakteristiky jako oblíbený politický kandidát. A demografická data jsou obvykle levnější než názorová či behaviorální, zvláště jsou-li zakoupena dohromady. Jednou z nevýhod demografických dat je fakt, že je poměrně obtížné získat přesná data pro jednotlivce. Nejsou-li data poskytnuta na výrobek nebo službu, mnoho lidí odmítá tento druh informací poskytovat nebo poskytují záměrně nepravdivé informace.

2.1.2 Behaviorální data

Vyjadřují míru akce nebo chování. Behaviorální data jsou obecně typem dat poskytujícím nejlepší prediktivní sílu. V závislosti na odvětví mohou být jejich součástí prvky jako prodaná množství, typy a data nákupů, data a výše plateb, činnost zákaznických služeb, pojišťovací nároky, chování při krachu a podobně. Jiným typem behaviorálních dat jsou aktivity na webových serverech. Firemní web lze navrhnout tak, aby zachycoval prodeje, jednotlivá klepnutí uživatele nebo přesnou cestu procházení každého návštěvníka webem.

Behaviorální data obvykle plní úlohu předpovědi budoucího vývoje lépe než jiné typy dat. Je však obvykle také složitější a dražší taková data z vnějšího zdroje získat.

2.1.3 Psychografická data

Psychografická data jsou charakterizována názory, životním stylem či osobními hodnotami. Tento typ dat je tradičně spojován s výzkumem trhu a získává se hlavně prostřednictvím šetření, výzkumů mínění a zájmových skupin. Lze je také odvodit z nákupního chování. Díky

zostřené konkurenci se tento typ dat pro zlepšené cílené modelování a analýzy integruje do zákaznických databází.

Psychografická data přinášejí do prediktivního modelování další rozměr. Firmám, které vymáčkly ze svých demografických a behaviorálních dat všechnu prediktivní schopnost (informaci), nabízejí psychografická data určité další možnosti. Jsou také užitečná při určování životního stupně zákazníka či potenciálního zákazníka. To vytváří nové možnosti pro vývoj výrobků a služeb ve spojení s životními událostmi, jakými jsou sňatek, narození dítěte, vysokoškolské studium a důchod.

Největší nevýhodou psychografických dat je, že vyjadřují zamýšlené chování, které může vysoce, částečně nebo jen okrajově korelovat se skutečným chováním. Data lze získat pomocí různých šetření nebo zájmových skupina a aplikovat je na větší skupiny lidí na základě segmentace nebo jiné statistické metody. Jestliže jsou na data aplikovány tyto metody, doporučuje se, aby byla otestována existence korelace.

2.2 Zdroje dat

Data pro modelování lze získat z mnoha zdrojů. Tyto zdroje spadají do jedná ze dvou kategorií: interní a externí. Interní (vnitřní) zdroje jsou ty, které vznikají prostřednictvím aktivit firmy jako záznamy o zákaznících, firemní web, záznamy z poštovních či telefonních kampaní nebo databáze či datové sklady, které jsou přímo určeny k uchovávání firemních dat. Externí (vnější) zdroje mívají firmy jako úvěrové kanceláře, obchodníci a kompilátoři seznamů a společnosti s rozsáhlými databázemi zákazníků jako vydavatelé a katalogoví prodejci.

2.2.1 Interní zdroje

Interní zdroje dat jsou ty, které jsou udržovány uvnitř firmy nebo organizace. Jde často o data s nejvyšší prediktivní schopností pro modelování, protože reprezentují informace, které jsou specifické pro výrobky a služby dané firmy. Typickými zdroji interních dat jsou databáze zákazníků, databáze provedených transakcí, databáze historie nabídek, pásky se záznamy požadavků a datové sklady.

2.2.2 Externí zdroje

Mnohé firmy se ocitají pod tlakem zvýšit zisky buď získáním nových zákazníků, nebo zvýšením tržeb od stávajících zákazníků. Obě tyto iniciativy lze vylepšit využitím externích zdrojů.

Mezi externí zdroje se řadí obvykle prodejci a kompilátoři seznamů. Jak byste čekali, prodejci seznamů jsou firmy, které prodávají seznamy osob. Jen málo z těchto firem však má prodej seznamů jako svůj výhradní předmět činnosti. Mnohé z nich se zabývají v první řadě prodejem prostřednictvím časopisů nebo katalogů a prodej seznamů a osob bývá jejich vedlejší činností. Podle druhu činnosti obvykle shromažďují a prodávají jména, adresy a telefonní čísla společně s demografickými, behaviorálními či psychografickými údaji. Někdy také provádějí „očistu" seznamů nebo jejich pročištění, aby zvýšili jejich hodnotu. Mnohé z těchto firem prodávají své seznamy prostřednictvím kompilátorů a brokerů seznamů.

Kompilátoři seznamů jsou firmy, které prodávají různé seznamy, z nichž některé jsou založeny na jediném seznamu a jiné jsou kompilovány z několika různých databází. Některé firmy vycházejí z podkladů, jako je telefonní seznam nebo registrační data z řidičských průkazů. Pak nakupují seznamy, vzájemně je slučují a doplňují chybějící údaje. Mnohé z těchto firem provádí vlastní výzkumy, aby zdokonalily přesnost svých seznamů.

Existuje také řada firem, které prodávají seznamy jmen spolu a kontaktními údaji a osobními charakteristikami. Některé z nich se specializují na určité typy dat. Je dobře známo, že úvěrové kanceláře prodávají data vystihující chování lidí při používání kreditních karet. Pomáhají finančním institucím při získávání a poskytování informací ohledně kreditního chování jejich členů. Firem prodávajících seznamy existují doslova stovky, od úzce specifických až po celostátní pokrytí.

2.3 Klasifikace dat

Existují dvě třídy dat kvalitativní a kvantitativní

2.3.1 Kvalitativní data

V kvalitativních datech se odlišují proměnné pomocí popisných pojmů. Například pohlaví se všeobecně klasifikuje jako „M" jako male, muž, a „F" jako female, žena.) Kvalitativní data lze použít pro segmentaci a klasifikaci.

2.3.2 Kvantitativní data

Kvantitativní data jsou charakteristická číselnými hodnotami. Pohlaví by mohlo být rovněž kvantitativní povahy, kdyby byla stanovena hodnotami 1 a 2 tak, že 1 = „M" čili muž a 2 = „Z", tedy žena. Kvantitativní data se používají k vytváření prediktivních modelů. Rozlišujeme čtyři typy kvantitativních dat.

2.3.2.1 Nominální data

Jsou číselná data, která reprezentují kategorie neboli atributy. Číselné hodnoty pro pohlaví (1 a 2) by byla nominálními datovými hodnotami. Důležitou vlastností nominálních dat je to, že nemají relativní význam. Například, i když muž = 1 a žena = 2, není relativní hodnota toho, že je někdo žena, dvakrát větší ani vůbec větší, než že je někdo muž. Pro modelování by měla být nominální proměnná s pouze dvěma hodnotami kódována hodnotami 0 a 1.

2.3.2.2 Ordinální data

Jsou číselná data, která představují kategorie, jež mají relativní význam. Lze je použít k hodnocení síly nebo důležitosti. Například firma zabývající se prodejem seznamů přiřadí určité proměnné hodnoty od 1 do 5 pro vyjádření finančního rizika. Hodnota 1, vyjadřující vždy včasné splácení, bude považována za nízké riziko. Hodnota 5, vyjadřující bankrot, bude znamenat vysoké riziko. Hodnoty 2, 3, 4 budou vyjadřovat různé úrovně rizika mezi nízkým a vysokým rizikem. Potenciální zákazník s hodnotou rizika 5 bude zcela jistě rizikovější než zákazník s hodnocením 1. Nebude však pětkrát riskantnější. A rozdíl 5 – 1= 4 mezi jejich hodnoceními nemá žádný význam.

2.3.2.3 Intervalová data

Jsou číselná data, která mají relativní význam a nemají nulový bod. V jejich případě jsou sčítání a odčítání smysluplnými operacemi. Například řada finančních institucí používá hodnocení rizika s daleko jemnější definicí než pouhé hodnoty 1 až 5, jak tomu bylo v předchozím příkladě. Typický rozsah bývá od 300 do 800. Pak je možné porovnávat hodnocení měřením rozdílů.

2.3.2.4 Spojitá data

Jsou nejčastějším typem dat používaných při vytváření prediktivních modelů. Se spojitými daty lze provádět všechny základní aritmetické operace včetně sčítání, odčítání, násobení a dělení. Většina obchodních údajů, jako jsou tržby, zůstatky či minuty, jsou spojitá data.

2.4 Výběr dat pro modelování

Výběr vhodných dat pro vytvoření cíleného modelu vyžaduje důkladné porozumění trhu a vlastním cílům. I když i nástroje jsou důležité, data slouží jako rámec nebo informační základna. Model je jen tak kvalitní a relevantní, jaká jsou jeho zdrojová data.

Zabezpečení dat může spočívat v extrakci dat ze stávajících zdrojů nebo může znamenat vytvoření vlastních zdrojů. Odpovídající výběr dat pro vytvoření a validaci cíleného modelu je klíčem ke zdárnému modelu.

2.4.1 Data pro získání zákazníků

Nejlepší volbou pro cílené modelování jsou data z některé předchozí kampaně. To platí, ať už se tato kampaň přesně shoduje s výrobkem či službou, kterou modelujeme, nebo ne. Kampaně vytvořené vaší firmou budou jistě kreativní a budou vhodně odrážet identifikaci zákazníků s vaší značkou. To může mít příznivý vliv na věrohodnost zákazníka.

Nemáte-li data z předchozích kampaní k dispozici, je dalším vhodným postupem vytvořit model sloužící k výpočtu tendence zákazníků k nákupu určitého typu výrobku. Tato modelovací metoda využívá dat z externího zdroje a z něj vytváří model, který cílí na výrobek či službu podobnou vašemu hlavnímu cíli.

Stále více firem navazuje spolupráci s jinými firmami, aby sdílely datové zdroje a zvýšily tak svůj zisk. Kreditní banky vytvářejí partnerství s leteckými společnostmi, univerzitami, kluby, maloobchodními prodejci a řadou jiných společností. Telekomunikační společnosti vytvářejí aliance rovněž s leteckými společnostmi, pojišťovnami apod. Jednou z hlavních výhod je přístup k osobním datům, která lze využít při vytváření cílených modelů. [6]

3 Dolování dat

Složitější požadavky na analýzu jsou řešeny prostřednictvím technologii dolování dat (data miningu). Dolování dat na základě určitého předpokladu umožňuje vyhledat ve velkém objemu dat souvislosti a vzájemné vztahy, které nebyly předem známy. Tím, že umí vyhledávat souvislosti v datech, které nebyly dopředu známy, se dolování dat liší od jiných metod počítačem zpracovávaných datových analýz.

Dolování dat by mělo mít vždy za cíl řešení konkrétního obchodního problému nebo nalezení cesty k vylepšení procesu. Cíl musí být předem definován a na jeho základě by se měla připravovat data. Dolování dat je hledáním skrytých souvislostí, procesem výběru, prohledávání a modelování ve velkých objemech dat. Slouží k odhalení dříve neznámých vztahů mezi daty.

Data musí být samozřejmě očištěna od chyb, úplná a formáty z různých systémů musí být sjednoceny. To zajistí jejich uložení v datovém skladu.

Spektrum metod, které se využívají při budování modelu dolování dat, je velmi rozsáhlé. O žádném modelu nelze říci, že je univerzální, a nejlepších výsledků se dosahuje kombinací různých přístupů.

Dnes užívanými metodami dolování dat jsou například:

- odhady hodnot vysvětlované proměnné (regresní analýza, neuronové sítě);
- klasifikace (diskriminační analýza, logistická regresní analýza, rozhodovací stromy, neuronové sítě);
- segmentace – shlukování (shluková analýza, genetické algoritmy, neuronové shlukování – Kohonenovy mapy);
- analýza vztahů (asociační algoritmus pro odvozování pravidel typu „if X then Y“);
- predikce v časových řadách (Boxova-Jenkinsonova metoda, neuronové sítě);
- detekce odchylek.

Některé přístupy jsou založeny na přesně popsaném matematickém modelu a aplikace na konkrétní data sestává z testování hypotéz a výpočtu neznámých koeficientů. Druhá skupina modelů mění svou strukturu dynamicky, na základě dat, která zpracovávají. Užívané metody dolování dat jsou často bedlivě střeženým know-how SW firem. [1]

Mohlo by se zdát, že data mining je univerzální metoda, ale není to tak. Někdy můžeme na základě náhodně vybraných vstupů získat cenné informace, tedy přijdeme k nim jako „slepí k houslím“, jindy může být výsledek data miningu triviální, v praxi nevyužitelný. Nemá smysl při procvičování data miningu aplikovat data miningový model na databázi naplněnou náhodnými údaji. Kde žádné údaje nejsou, nemůžeme samozřejmě nic vydolovat. I při reálných údajích musí být tyto údaje kvalitní, tedy kvalitně připravené.

Může nastat problém, pokud v našich vstupních údajích nemáme podchyceny některý důležitý údaj, může být výsledek nesprávný. Příkladem je zkoumání denního prodeje piva bez zahrnutí vnější teploty a podobně.

Úlohu data miningu chápeme jako prostředek k získávání informací pro podporu rozhodování. Samotné rozhodování musí provést příslušný odpovědný pracovník, je to tedy prostředek k poskytnutí kvalitních vstupů do procesu rozhodování. Velkou úlohu managementu je vypracování a definování vizí, koncepcí, cílů a také vyslovování hypotéz. Data mining pak může posloužit jako nástroj pro ověření jednotlivých hypotéz, tedy zda je management nakonec zamítne, nebo nezamítne.

3.1 Datová kvalita

Datová kvalita je jednou ze základních vlastností datových skladů a operativních úložišť dat. Datová kvalita není jejich vlastností automatickou, je však v dnešní době vlastností nezbytnou.

Datovou kvalitu lze definovat různými způsoby, v tomto případě zvolme jednoduché vymezení – kvalitní data jsou taková, která odpovídají realitě, jsou úplná a konzistentní.

Pokud chceme pracovat s kvalitními daty, musíme zajistit jejich pět základních vlastností:

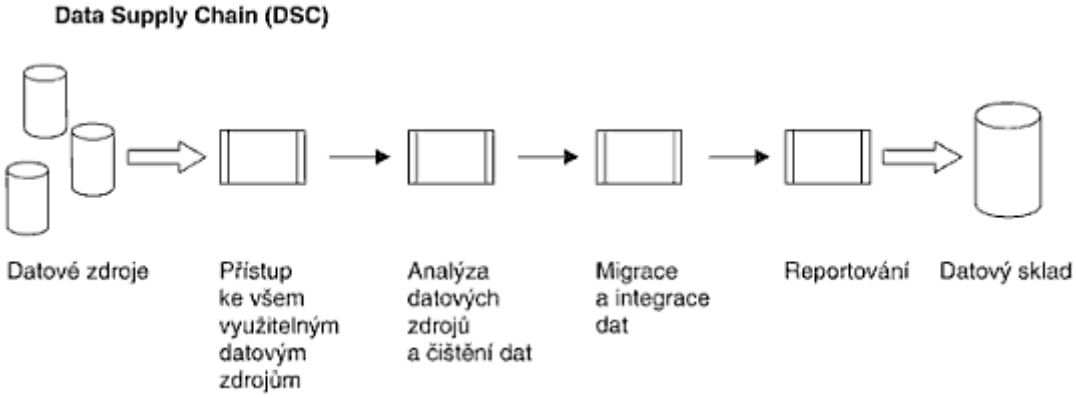
- úplnost-potřeba identifikovat a ošetřit data, která chybí nebo jsou nepoužitelná;
- soulad-všechna data by měla odpovídat požadovanému formátu;
- konzistenci-žádná data nesmějí obsahovat hodnoty, jež reprezentují konfliktní informace;
- unikátnost-pokud existují duplicitní záznamy, musí být odstraněny;
- integritu-všechna data by měla obsahovat veškeré definované vztahy vůči ostatním datům.

Problematika datové kvality získala na významu v době budování datových skladů, kdy docházelo ke zjištění, že data zpracovávaná z různých zdrojů ve větší nebo menší míře neodpovídají realitě, případně obsahují konfliktní hodnoty.

Datová kvalita je v současné době většinou dodavatelů informačních technologií řešena jako množina dvou procesů:

- analýzy kvality dat (data profiling);
- korekci a opravy dat (data cleansing).

Tyto procesy se pravidelně opakují a promítají se do prostředí datového skladu, hlavně ETL procesu. Vzniká tak komplexní proces nazývaný Data Supply Chain (DSC). DSC je automatizovaný proces, jehož prostřednictvím jsou data z provozních systémů přenášená do datového skladu (viz. obr. 1). V jeho první fázi jsou extrahována data z datových zdrojů ať již interních nebo externích. V druhé fázi je prováděná analýza datových zdrojů a čištění dat.



obr. 1 Řetězec dodávky dat

3.2 Datový sklad

Datový sklad je centrální úložiště různorodých dat firmy, které obsahuje data v databázi, ale také nástroje pro výběr a filtrování dat a jejich analýzu. Všechny informace získané v datovém skladu lze podle potřeby jednoduchým, uživatelsky přívětivým způsobem prezentovat. Současně musí být všechna data zabezpečena proti případnému zneužití. To je nutné z několika důvodů. Jedním z nich jsou náklady, které firmy na vytvoření datových skladů vydávají. Náklady se jim později vrací při využívání informací získávaných prostřednictvím výstupů datových skladů (datových tržišť). Pro danou firmu tedy není žádoucí, aby se datového skladu zmocnila např. konkurenční firma.

Dalším důvodem může být platnost některých zákonů, např. na ochranu osobnosti, protože datový sklad může obsahovat také velmi citlivé údaje.

Přístup k údajům v datovém skladu je tedy závislý na uživatelském oprávnění, které vyplývá z potřeb a požadované spolehlivosti konkrétních pracovníků. Celý proces tvorby a udržování datových skladů, ale také jejich využívání se označuje pojmem business inteligence. [2]

3.2.1 Podnikový sklad

Podnikový sklad (enterprice warehouse) sbírá všechny informace o subjektech, které obklopují celou organizaci. Provádí integraci celopodnikových dat pocházejících obvykle z jednoho nebo více provozních systémů nebo od externího poskytovatele informací. Tato data zasahují do řady oborů. Obvykle obsahují jak hodnoty detailní, tak i sumarizované. Jeho velikost se může pohybovat od několika gigabyte až po stovky terabyte. Tento typ skladů bývá implementován na mainframy, Unixové superservery nebo na paralelní platformu. Vyžadují rozsáhlé modelování a jejich návrh a vytvoření může trvat několik let.

3.2.2 Datové tržiště

Datové tržiště (data mart) obsahuje pouze podmnožinu celopodnikových dat, která je určená pro specifickou skupinu uživatelů. Rozsah dat je omezen na určité vybrané subjekty. Např. v marketingovém tržišti jsou obsaženy informace týkající se zákazníků, zboží a prodejů. Tyto hodnoty bývají sumarizovány.

Datová tržiště jsou implementována na levnější servery s Unixovým nebo Windows/NT jádrem a jejich tvorba se pohybuje v řádu týdnů. Podle zdroje získávání dat rozlišujeme data marty na „nezávislé“ (získávají data z provozních systémů nebo z externích informačních zdrojů) a „závislé“ (data jsou jim dodávána z podnikového datového skladu).

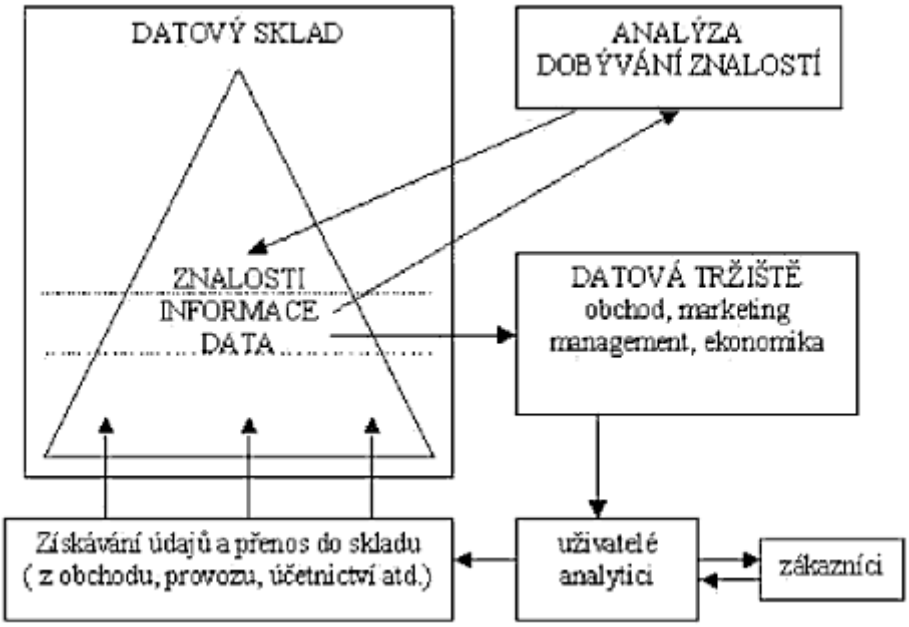
3.2.3 Virtuální sklad

Virtuální sklad (virtual warehouse) je sadou náhledů na provozní databáze. Pro efektivnější provádění dotazů jsou některé náhledy na sumarizace provedeny před vznikem vlastního požadavku a uloženy. Virtuální sklad je snadné vytvořit, ale vyžaduje dodatečné kapacity na provozních serverech. [3]

3.3 Business intelligence

Pojem business intelligence označuje proces transformace dat a převod těchto informací na znalosti, sloužící k podpoře podnikání.

Architekturu business intelligence, tj. různé souvislosti při využívání datových skladů ukazuje obr.2. Základem je sběr a ukládání obrovského množství údajů, získaných například z obchodních transakcí, z provozu, účetnictví apod. Tyto vstupní údaje jsou zprvu neroztříděné, nefiltrované, tzn. nezpracované.



obr. 2 Jednoduché schéma business intelligence

V datovém skladu dochází k ukládání dat a jejich následnému zpracování. Filtrováním se odstraní nadbytečné údaje, které se třídí, ověřuje se jejich správnost a zařazují se do skladu podle různých potřebných kritérií. Tím se počet údajů sice sníží, ale zbylé informace mají podstatně vyšší vypovídací hodnotu. Mohou být následně prostřednictvím datových tržišť a díky OLAP analýze využívány při řízení firmy, při obchodních jednáních a pro potřeby marketingu.

Pomocí analýzy a dobývání znalostí z datového skladu lze získat i skryté informace, které mohou být využity pro konkrétní účely. Například v marketingu můžeme tímto způsobem získat potřebné podklady pro rozhodnutí, které zákazníky je vhodné oslovit s nabídkou nového produktu.[2]

3.4 Analýza OLAP

Termín OLAP zavedl Dr. E. F. Cood na popsání technologie, která by pomohla překlenout mezery mezi využitím osobních počítačů a řízením podnikových dat. Pro OLAP existuje více, například:

OLAP je volně definovaný řád principů, které poskytují dimenzionální rámec pro podporu rozhodování.

Pojem OLAP se poměrně často zaměňuje s jiným pojmem DSS (Decision Support Systems) – systémy na podporu rozhodování. Tyto systémy umožňují pracovníkům přijímacím rozhodnutí přístup k údajům potřebným na „tvorbu" takových rozhodnutí.

3.4.1 Relační databázový model

Údaje jsou uloženy v dvoudimenzionálních tabulkách. Každý řádek v tabulce obsahuje data, která jsou zpravidla obrazem reálného světa, tedy data, která se vztahují k nějaké věci nebo její části. Sloupce dvoudimenzionálních databázových tabulek obsahují údaje týkající se atributů.

3.4.2 Multidimenzionální databázový model

Datový model multidimenzionální databáze je možné zobrazit jako vícerozměrnou krychli. Tato krychle je vlastně ekvivalent tabulky v relační databázi. Každá krychle má několik dimenzí (ekvivalent indexových poli v relačních tabulkách). Nejlépe si dokážeme představit klasickou trojrozměrnou krychli, ale počet dimenzí v reálných multidimenzionálních databázích je zpravidla větší.

S rostoucím počtem rozměrů multidimenzionální databáze velmi rychle rostou i požadavky na úložnou kapacitu. Ale ne na všech průsečících dimenzí se vždy nacházejí údaje. Takovou krychli nazýváme i řídkou krychlí. V praxi se u multidimenzionálních databází používají různé technologie na kompresi objemu použitého diskového prostoru.

3.4.2.1 Multidimenzionální OLAP (MOLAP)

Pro multidimenzionální online analytické zpracování (MOLAP) se získávají data buď z datového skladu, nebo z operačních zdrojů. Mechanismus MOLAP potom uloží analytická data ve vlastních datových strukturách a sumářích. Během tohoto procesu se spočítá tolik předběžných výsledků, kolik je z technického a časového hlediska možné. Údaje v úložišti

typu MOLAP se tedy budou ukládat jako dopředu vypočítaná pole. Hodnoty dat i indexů se uchovávají v jednotlivých polích multidimenzionální databáze. Databáze je organizována tak, aby umožnila rychlé získávání příslušných údajů z více dimenzí. Část údajů se může zavést ze serveru ke klientovi, což umožňuje rychlé analýzy bez velkého zatížení sítě.

3.4.2.2 Relační databázový OLAP (ROLAP)

Relační online analytické zpracování údajů (ROLAP) získává údaje pro analýzy z relačního datového skladu. Tyto údaje z relačních databází se po zpracování předkládají uživateli jako multidimenzionální pohled. Data a metadata se v úložišti ROLAP ukládají jako záznamy v relační databázi. Server OLAP dynamicky používá tato metadata na generování příkazů SQL, které jsou potřebné na získávání dat požadovaných uživateli. U tohoto způsobu zůstávají data uložena v relačních databázích, takže nevzniká problém s redundancí.

3.4.2.3 Hybridní OLAP (HOLAP)

Hybridní OLAP je kombinací úložišť MOLAP a ROLAP, přičemž se využívají výhody jednotlivých typů úložišť a do značné míry se eliminují nevýhody. Údaje zůstávají v relačních databázích a spočítané agregace se ukládají do multidimenzionálních struktur. Při dotazování se údaje vybírají do multidimenzionální paměti cache. U hybridního řešení relační databáze ukládá množství detailních dat a multidimenzionální model ukládá sumární data. [10]

3.5 Databáze

Databáze je obvykle rozsáhlý počítačový soubor, obsahující seznam informací (jména, adresy a jiná data pro zákazníky, zaměstnance atd.). Uživatelé databáze používají speciální software, aby v ní našli informace a aktualizovali data. [9]

3.6 Relační databáze

Databázový systém, který se také jmenoval databázový systém (DBMS), se skládá z kolekce vzájemných údajů, známých jako databáze, sadou zvládnutých softwarových programů a přístupem dat. Programy zahrnují postupy pro definici databázových struktur, pro ukládání dat, pro souběžné, sdílené, nebo distribuci přístupu k datům pro zajištění konzistence a zabezpečení uložených informací, a to navzdory zhroucení systému nebo pokusy o neoprávněný přístup.

Relační databáze je soubor tabulek, z nichž každý je přiřazen jedinečný název. Každá tabulka se skládá ze sady atributů (sloupců nebo polí) a obvykle ukládá velký soubor n-tic

(záznamů nebo řádků). Každá n-tice v relační tabulce reprezentuje objekt označen jedinečným klíčem a popsané souborem hodnot atributů. Sémantický datový model, jako subjekt-vztah (ER) datového modelu je často konstruován pro relační databáze. Data ER model reprezentuje databáze jako soubor entit a jejich vztahů. [4]

3.7 Zneužití dat

V dnešní době je trend koncentrovat a uchovávat velká množství dat v různých databázích. Tato data fyzicky nezabírají skoro žádné místo a v jednom bodě může být takových dat nashromážděno velké množství. Další rys vyplývající z možnosti ukládání dat je, že uložená data jsou neustále k dispozici a může je někdo dále kdykoliv zpracovat. A to jak jednotlivě, tak také hromadně. Data se mohou třídit a ukládat do různých databází, propojovat a mohou se tak v nich hledat různé souvislosti. A právě v kombinování jednotlivých údajů a v jejich spojování s dalšími informacemi je asi největší nebezpečí zneužití osobních dat.

Nikdo nikdy nebude vědět, kolik informací a jakého druhu je o jeho osobě kde uloženo. Z tohoto důvodu může dojít k situacím, kdy výběr vhodně zkombinovaných dat o určité osobě může vytvořit odraz jeho soukromí. Z informací, které se tak chytře spojí se stane zboží. Zboží, které bude mít svou určitou hodnotu a to nejen pro ty, kteří nás budou chtít kontaktovat a prodat nám svůj produkt nebo službu.

3.8 Historie data miningu

Když nahlédneme zpět do historie je jedna forma data minigu známa jako data dredging, tedy „bagrování dat". Tento obor byl považován za něco, co je pod úrovní dobrého výzkumníka. Pojem naznačoval, že výzkumník může skutečně prohledávat dat a bez jakýchkoli předběžných hypotéz. V poslední době se však této technice dostalo mnohem lepšího přijetí, zejména proto, že tato forma dolování dat vedla k objevení velmi cenných informací. V Americe, založené na korporátních společnostech, platí, že odkryje li proces informace vedoucí ke zvýšení zisků, je rychle přijat a získává respekt.

Jiná forma dolování dat začala získávat popularitu v marketingové aréně na konci osmdesátých let a počátkem devadesátých let dvacátého století. Několik bank specializovaných na kreditní karty spatřilo ve formě data miningu známe jako data modeling cestu k vystupňování akvizičních aktivit a vylepšení řízení rizika. Nevídaná akvizice a bezprecedentní růst se stal úrodnou půdou rozkvět disciplíny modelování dat. Úspěšné a profitabilní využití data modelingu otevřelo cestu uplatnění a zdokonalení těchto technik do

jiných odvětví. Dnes mezi obory pracující s technikami modelování dat patří pojišťovnictví, přepážkové i investiční bankovnictví, veřejné služby, telekomunikace, zásilkové služby, energetika, maloobchod, cestovní ruch, zábava, farmaceutický průmysl, ale celý seznam by byl mnohem delší. [6]

4 Metody dolování dat

Jak již bylo v předchozí kapitole napsáno existuje řada metod dolování dat. Proto bych chtěl v této práci představit alespoň několik z nich.

4.1 Neuronové sítě

Neuronové sítě jsou systémy, které provádějí rozpoznání vzorů v přijatých vstupech na základě modelů, jak zpracovávají informace neurony savců. Neurony jsou nervové buňky vzájemně hustě propojené a stýkající se v synapsích, malých výběžcích mezi jednotlivými neurony. Rovněž paralelně pracují s ostatními neurony v libovolné úrovni mozkové struktury. Neuronové sítě jsou sadou matematických modelů imitujících učební proces neuronů, přiřazujících různé váhy spojením mezi vnitřními prvky neuronové sítě podobným způsobem, jako jsou předávány elektrické potenciály na neuronové synaptické spoje v závislosti na frekvenci jejich vzruchů. Čím častěji je sousední neuron podrážděn, tím větší elektrický potenciál se objeví na jeho synapsích neuronů, které reagují na tento vzor. V neuronových sítích ty prvky, které přijaly vstup od sousedních prvků, získávají větší váhu.

Ačkoliv je to komplikované a stále poněkud záhadné, přístup neuronových sítí může být aplikován na problém rozpoznání vzorů v širokém rozsahu, včetně detekce narušení. Elegance neuronových sítí v detekci narušení spočívá v tom, že veškerá pravidla nebo signatury jsou zde nadbytečné. Jednoduše zahájíte naplněním vstupu – daty vztahujícími se k síťovým nebo uzlovým událostem, do neuronové sítě a ta se postarají o zbytek. Proto jsou neuronové sítě velmi vhodné a pohotově připravené ke sběru nových útočných vzorů, ačkoli je samozřejmě zapotřebí jistého času, než se to naučí. Přístup spojený s neuronovými sítěmi se těší přízni již delší dobu, a jestliže je něco pravděpodobné, pak to, že, jak se bude zmenšovat spolehlivost a význam signatur, se stanou široce používanými a spolehlivými v detekci narušení. [5]

4.2 Rozhodovací stromy

Rozhodovací stromy jsou analytické nástroje sloužící k nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně.

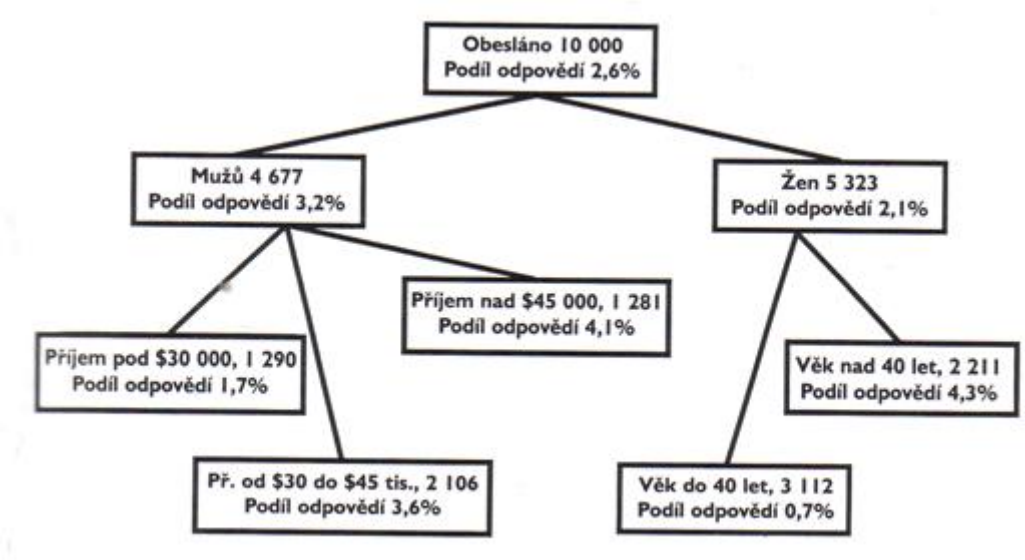
Cílem je určit takové proměnné, které dokážou záznamy rozdělit a snižují tak nejistotu. Problémem může být určení, na kolik „větví“ se má dělit každá proměnná. Pokud záznamy rozdělíme podle proměnné do příliš mnoha skupin, může nastat situace, kdy do každé z těchto skupin přísluší pouze několik málo záznamů a nelze tak vyvodit žádná rozhodovací pravidla.

Rozhodovací stromy jsou vhodné pro úlohy, ve kterých má být provedena klasifikace nebo předpověď. Užitečné jsou v oblastech, ve kterých můžeme hodnoty proměnných rozdělit do relativně malého počtu skupin. Na druhou stranu nejsou vhodné pro případy, kdy je úkolem předpovědění kvantitativních hodnot.

Strom se skládá z uzlů. Uzel na nejvyšší úrovni je označován pojmem kořenový. Vnitřní uzly představují testy jednotlivých atributů (kořenový uzel je rovněž testem). Větví nazýváme možný výsledek testu. Externí uzly označované jako listy reprezentují jednotlivé třídy. [3]

Klasifikační stromy jsou „vytvářeny“ řadou kroků a pravidel, které nabízejí vysokou flexibilitu. Na obrázku 3 strom rozlišuje mezi respondenty a nerespondenty. Vrcholový uzel představuje úspěšnost celé kampaně. Propagační materiál byl zaslán 10 000 lidí s odezvou 2,6 %. První rozdělení proběhne na základě pohlaví. Z toho vyplývá, že největší rozdíl mezi těmi, kdo odpoví, a těmi, kdo neodpoví, je dán pohlavím. Vidíme, že muži na nabídku odpovídají mnohem více (3,2 %) než ženy (2,1%). Pokud se zastavíme po prvním dělení, budeme považovat muže za lepší cílovou skupinu. Naším cílem však je najít skupiny odpovídajících mezi oběma pohlavími. Při dalším dělení zvažujeme obě tyto skupiny či uzly zvlášť.

Dělícím kritériem na druhé úrovni je u mužů příjem. Z toho vyplývá, že se mezi těmi, kdo odpověděli, a těmi, kdo nikoli, nejvíce liší jejich příjem. U žen je největší rozdíl mezi věkovými skupinami. Skupiny s nejvyšší mírou reakce lze najít velice snadno. Řekněme, že se management rozhodne oslovit pouze ty skupiny, u nichž je míra reakce vyšší než 3,5 %. Zásilky tedy budou zaslány mužům, kteří vydělávají více než 30 000 dolarů ročně, a ženám nad 40 let.



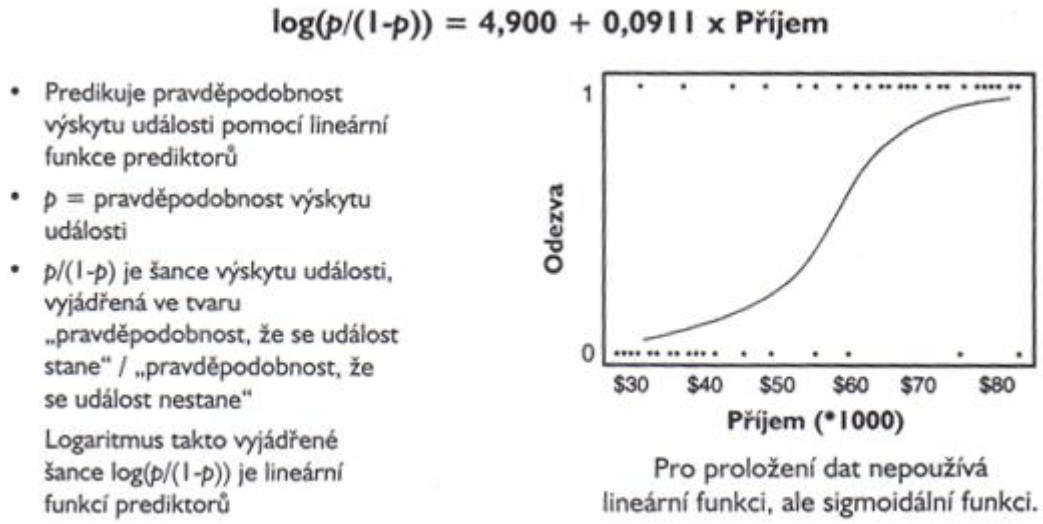
obr. 3. Klasifikační strom odpovědí

4.3 Logistická regrese

Logistická regrese je velmi podobná lineární regresi. Hlavní rozdíl spočívá v tom, že závislá proměnná není spojitá; je diskrétní neboli kategorická. Tím se stává velmi užitečná v marketingu, protože se často snažíme předvídat diskrétní akci, například odezvu na nabídku nebo nesplacení půjčky.

Logistickou regresi lze používat k predikování výsledků dvou nebo více úrovní. Při vytváření cílených modelů pro marketing však bývá výsledek obvykle dvouúrovňový. Aby se dala využít regrese, transformuje se závislá proměnná na spojitou hodnotu, která je funkcí pravděpodobnosti výskytu události.

Na obrázku 4 je zobrazen graf vztahu mezi odezvou (0/1) a příjmem v dolarech. Cílem je predikovat pravděpodobnost odezvy na katalog s nabídkou prestižních dárků na základě příjmu zákazníka. Datové body nabývají pro odezvu hodnot 0 nebo 1. A na ose příjmu jsou hodnoty 0 odezvy soustředěny okolo nízkých hodnot příjmů. Sigmoidální funkce – křivka ve tvaru S – je vytvořena zprůměrováním nul a jedniček pro každou hodnotu příjmu. Zákaznici s vysokým příjmem reagují na nabídku ve vyšší míře než zákazníci s nízkým příjmem.



obr. 4 Logistická regrese

Zpracování modelu je následující:

1. Pro každou hodnotu příjmu se vypočítá pravděpodobnost *(p)* zprůměrováním hodnot odezvy.
2. Pro každou hodnotu příjmu se vypočítají šance pomocí vzorce *p(1-p),* kde *p* je pravděpodobnost výskytu události.
3. Při finální transformaci se počítá logaritmus šance: *log(p/(1-p))*. [6]

4.4 Kohonenovy mapy

Pro kategoriální klasifikaci velké kolekce dokumentu se používají Kohonenovy samoorganizované mapy (Kohonen Self-Organizing Maps (SOM)). Tato metoda provádí projekci z vysocerozměrného prostoru dokumentu do prostoru s nižší dimenzí – většinou dvourozměrné mřížky. Důležitou vlastností projekce je částečné zachování topologie, tj. blízké body v původním prostoru jsou blízko i v mřížce.

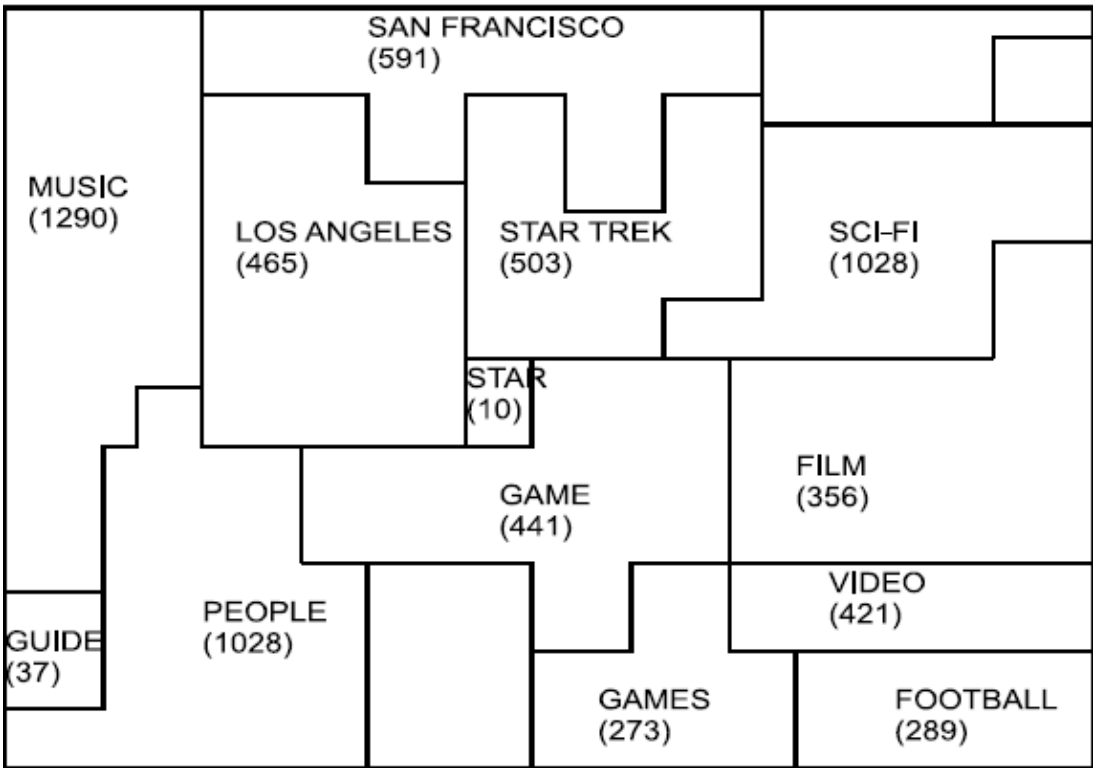
V případě IR, metoda SOM „vyrábí" grafy podobnosti (similarity graphs) mezi dokumenty. Na vstupu (ve vstupní vrstvě sítě) jsou vektory vlastností jednotlivých dokumentu. Metoda SOM v průběhu zpracování dokumentu vytváří v Kohonenově mapě shluky, ke kterým jsou přiřazovány zpracovávané dokumenty. Toto shlukování je prováděno na základě zvýšené excitace vítězného neuronu a jeho sousedu.

Dílčím problémem je velikost a reprezentace vektoru vlastností dokumentu, který vstupuje do neuronové sítě. V naivním vektorovém modelu je to například vážený histogram slov, která se v daném dokumentu nacházejí. Takováto reprezentace ale není vhodná, protože

velikost vektoru (a tudíž i dimenze původního prostoru) je v řádu desetitisíců, což pro Kohonenovu mapu není z hlediska efektivity ideální. Pro tyto účely se používají různé předzpracující techniky redukce dimenze vektoru vlastností dokumentu:

- Latentně-sémantická indexace (LSI). Tato maticová metoda poměrně účinně redukuje dimenzi tak, že pomocí SVD (singular-value decomposition) rozkladu je vytvořen předem určený (malý) počet důležitých faktorů, které dohromady tvoří význam jednotlivých dokumentů. Do Kohonenovy mapy pak vstupují vektory s dimenzí rovnou počtu těchto faktorů.
- Náhodná projekce histogramů. Experimentálně se ukázalo, že prosté promítnutí vektoru histogramu do prostoru nižší dimenze vede k dobrým výsledkům bez ztráty signifikantních informací odlišujících dokumenty.
- Mapy pro slovní kategorie. Elegantním řešením je opět použití metody SOM pro redukci dimenze vstupního vektoru. Speciální „slovní“ Kohonenova mapa nyní shlukuje slova do slovních kategorií, přičemž slučuje synonyma, různé tvary stejných slov a neslova do významových kategorií. Do původní „dokumentové“ Kohonenovy mapy pak vstupují vektory s dimenzí rovnou počtu slovních kategorií.

Na obrázku 5 je vytvořená Kohonenová mapa a shluky podobných dokumentů.



obr. 5 Shluky dokumentů jako Kohonenova mapa

Dotaz do hotové mapy vrátí jako odpověď všechny dokumenty do určité vzdálenosti v nalezeném shluku. [7]

5 Metodologie CRISP-DM

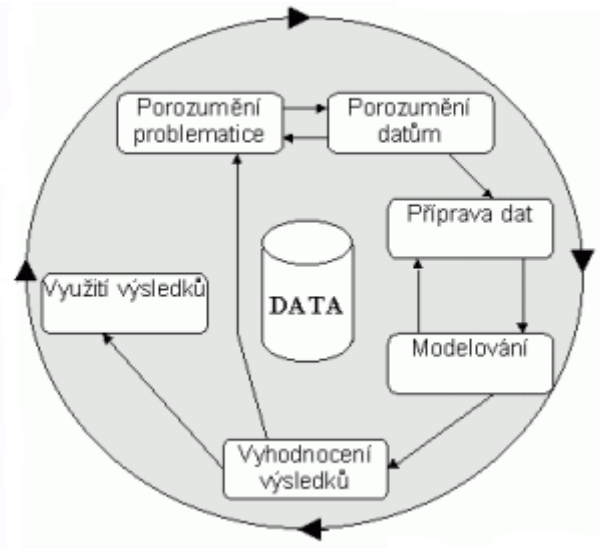
Data miningový projekt je proces, který vyžaduje značné množství zdrojů. Od lidských, přes hmotné a datové až po softwarové. Společným jmenovatelem jim jsou peněžní prostředky. Jedním ze způsobů, jak šetřit tyto prostředky, je provádět projekty standardizovaným postupem.

Za tímto účelem byly vytvořeny různé metodologie popisující efektivní postup při projektech.

Zkratka CRISP-DM znamená CRoss-Industry Standard Proces for Data Mining. CRISP je souhrnná data miningová metodologie. Její model nabízí návody krok po kroku, úkoly a cíle pro každou část celého procesu. CRISP-DM umožňuje provádět rozsáhlé data miningové projekty rychleji, efektivněji a méně nákladně prostřednictví osvědčených postupů. Model pomáhá vyhnout se běžných chybám.

Vývoj metodologie CRISP-DM byl zahájen jako projekt Evropské komise definující model standardního postupu při vytváření data miningových projektů. [3]

Životní cyklus projektu dobývání znalostí je podle metodologie CRISP-DM tvořen šesti fázemi (Obr. 6). Pořadí jednotlivých fází není pevně dáno. Výsledek dosažený v jedné fázi ovlivňuje volbu kroků následujících, často je třeba se k některým krokům a fázím vracet. Vnější kruh na obrázku symbolizuje cyklickou povahu procesu dobývání znalostí z databází jako takovou.



obr. 6 Metodologie CRISP-DM [8]

5.1 Porozumění problematice (definování cílů)

Vstupní fáze je zaměřena na definování cílů projektu a požadavků z obchodního hlediska. Poté na převedení znalostí na definici data miningového problému a předběžné navržení plánu, jak dosáhnout cílů.

Měl by být definován základní cíl projektu z podnikatelského hlediska. K základnímu cíli jsou obvykle připojeny ještě další otázky, na které by klient rád dostal odpověď. Přestože metodologie uvádí jako součást těchto úvah i otázku prostředí firmy a její obchodní situace na trhu, je možné v některých případech tuto oblast vynechat bez negativního vlivu na výsledek projektu. Jedná se o projekty, jejichž výstupy neovlivní přímo okolí organizace.

Při definování cíle je zapotřebí definovat rovněž kritéria (z podnikatelského hlediska) pro hodnocení úspěšnosti nebo užitečnosti výstupu projektu. Tato kritéria mohou mít dvě odlišné formy. Mohou být objektivně měřitelná nebo subjektivně vnímatelná.

Před započetím projektu by měly být známy všechny vstupy, které budou nutné či dostupné. Tyto vstupy zahrnují jak časové, finanční a hmotné prostředky (prostory, hardware, atd.), tak softwarové, lidské (obchodní experti, datoví specialisté, techničtí pracovníci, data miningoví pracovníci) a datové zdroje (neměnné extrakty, přístup k datovým skladům, provozním datům).

Dalším dokumentem, který se v této etapě vypracovává, je analýza přínosů a nákladů. Tento dokument je nezbytnou součástí všech projektů, neboť každá organizace požaduje odpovědi na otázky: „Kolik mě projekt bude stát?“ a „Kolik peněž mi přinese nebo umožní ušetřit?“.

Nutnou součástí této etapy je sestavení plánu projektu, ve kterém je popsán způsob dosažení cílů data mining. Měly by být stanoveny kroky, které musí být vykonány, společně s jejich trváním, požadovanými zdroji, vstupy, výstupy a závislostmi. Součástí plánu je rovněž analýza závislostí mezi časovým harmonogramem a riziky. Projektový plán obsahuje detailní plán pro každou fázi. Plán je dynamický, což znamená, že na konci každé fáze je kontrován a aktualizován.

5.2 Porozumění datům

Další částí projektu je získání dat nebo přístupu k datům, která jsou uvedena ve zdrojích. Tento výchozí sběr zahrnuje případně i nahrání dat, pokud je to nutné pro jejich pochopení. Všechny tyto operace by měly být popsány společně s metodami užitými k získání dat. Zaznamenány by rovněž měly být i problémy vzniklé během tohoto procesu a způsoby řešení pro případné použití v budoucnosti (při opakování stejného či podobného projektu).

Popsáním charakteristik dat, jako např. formátu, množství dat (počtu záznamů a polí v každé tabulce), popisu polí a dalších znaků, byl měla být zodpovězena otázka, zda-li data uspokojují podstatné požadavky.

Již v této části se provádí zběžný průzkum dat. Tato analýza se zaměřuje na data miningové otázky, které mohou být zodpovězeny použitím dotazů, vizualizací a reporty. To zahrnuje: rozdělení klíčových atributů (např. cílová vlastnost pro úlohu predikce), vazby mezi páry nebo malým počtem atributů, výsledky jednoduché agregace, vlastnosti významných podskupin, jednoduché statistické analýzy. Tyto analýzy se mohou zaměřit přímo na cíl data miningového projektu a sloužit tak pro formulaci hypotéz, nebo pouze přispívat k popisu dat. Pokud je to vhodné, mohou být součástí i grafy a diagramy, které vyjadřují datové charakteristiky nebo které jsou vodítkem k zajímavým podskupinám v datech.

5.3 Příprava dat

Tato fáze bývá obvykle jednou z nejnáročnějších, neboť data bývají často v různých formátech, v různých tabulkách, obsahují chybějící hodnoty, jiné atributy potřebné pro analýzu chybějí úplně, atd.

Musíme rozhodnout, která data budou použita pro analýzu. Kritérii jsou: souvislost s cíli data mining, kvalita a technické podmínky jako např. omezení objemu dat nebo typů. V procesu výběru dat je nutné vybírat jak atributy (sloupečky), tak i záznamy (řádky) v tabulce.

Někdy vybraný nástroj či analytická technika vyžaduje, aby data měla určitou kvalitu. Potom je zapotřebí, aby záznamy prošly tzv. „čištěním“, což může zahrnovat např. vložení vhodných (standardních) hodnot nebo náročnější techniky (určení chybějících dat modelováním). Tyto změny by měly být dokumentovány zároveň se zvážením vlivu na výsledky analýz.

Organizace obvykle neshromažďují data s myšlenkou, že s nimi bude následně proveden tento typ analýz. Proto ve struktuře, v které jsou záznamy uloženy, mohou chybět atributy bezpodmínečně nutné pro projekt. Potom musí být provedeny operace, jenž tyto nedostatky odstraní. Zahrnují např.: vytvoření odvozených atributů, úplně nových atributů nebo transformace hodnot stávajících atributů.

Informace potřebné pro analýzu bývají uloženy v několika tabulkách. S tímto způsobem uložení však nedokáže většina data miningových nástrojů pracovat, a tak je zapotřebí sloučit data z několika tabulek do jediné. Sloučením se rozumí spojení dvou či více tabulek, které obsahují rozdílné informace o stejném objektu. Sloučená data mohou rovněž zahrnovat agregace. Agregací se mají na mysli operace, kde se nová hodnota počítá sumarizací informací z několika záznamů a/nebo tabulek

Posledním krokem v přípravě dat je jejich naformátování. Formátovací transformace se týkají v prvé řadě syntaktických změn prováděných na datech, které nemění jejich význam. Některé nástroje mají požadavky na pořadí atributů, jako např. první pole musí být jedinečný identifikátor záznamu nebo poslední pole musí být označení, které má model určovat. Může být nutná i změna pořadí záznamů v datové sadě. Modelovací nástroj může vyžadovat, aby záznamy byly seřazeny podle hodnot výstupního atributu (kterým se označují záznamy při klasifikaci). Běžnou situací je, že záznamy v datové sadě jsou uspořádány určitým způsobem. Modelovací algoritmus však může vyžadovat, aby byly náhodně uspořádány. Např. pokud

používáme neuronové sítě, je všeobecně lepší, aby data byla předána síti v náhodném pořadí, ačkoliv některé nástroje tuto změnu pořadí provádějí automaticky bez zásahu uživatele. Navíc se ještě provádějí zcela syntaktické změny odpovídající specifickým požadavkům jednotlivých modelovacích nástrojů. Např. odstranění čárek z textových polí v souboru, kde byly čárky použity jako oddělovače; zkrácení všech hodnot na maximálně 32 znaků.

5.4 Modelování

Před vlastním sestavením modelu potřebujeme vytvořit postup nebo mechanismus, který bude testovat kvalitu a sílu (správnost) modelu. Např. při klasifikaci používáme běžně jako měřítko kvality data miningového modelu počet chybných klasifikací v procentním vyjádření. Proto obvykle rozdělujeme datovou sadu na sadu učící a testovou. Model je vytvářen na učící datové sadě a jeho kvalita je určována na testové sadě dat.

V průběhu samotného modelování je vytvářen jeden nebo více modelů. V používaných nástrojích bývá často množství parametrů, které mohou být různě měněny. Proto je nutné vždy důkladně zaznamenat všechny nastavené hodnoty. Dle CRISP bychom měli zaznamenat rovněž zdůvodnění, proč jsme vybrali zrovna tuto kombinaci nastavení. Zapisování všech nastavení, SQL dotazů apod. usnadňuje potom např. orientaci v datech nebo zabraňuje opakování některých operací.

Nezbytnou součástí této etapy je ocenění modelů. Analytik ohodnocuje modely podle hledisek, kterými jsou v prvé řadě kritéria pro přesnost definovaná v první fázi. Pokud je to možné, bere rovněž v úvahu obchodní cíle a kritéria hodnotící obchodní úspěšnost. Protože však je jeho pohled spíše techničtější, spojuje se později s obchodním analytikem a expertem v dané oblasti, aby interpretovali výsledky v obchodních souvislostech.

5.5 Vyhodnocení výsledků

Předešlé hodnotící kroky používali pro hodnocení takové faktory jako přesnost a obecná platnost modelu. Tento krok hodnotí úroveň s jakou model dosahuje obchodních cílů a snaží se určit, zda-li je přítomen nějaký důvod (obchodní), proč je tento model nedostatečný. Vytvořený model je možné ohodnotit tím způsobem, že jej užijeme na reálné situace a sledujeme jeho kvalitu. Je však nutné zvážit časové a rozpočtové podmínky, zda-li umožňují takovéto hodnocení.

Pokud je výsledný model označen jako schopným uspokojit obchodní potřeby, následuje důkladná revize celé data miningové úlohy a určuje se, zda-li nebyl přehlédnut nějaký

důležitý faktor či úkol. Tato revize rovněž zahrnuje ujištění o kvalitě (o správném sestavení modelu; o použití atributů, které budou dostupné i pro budoucí analýzy).

S ohledem na výsledky hodnocení a revize procesu se rozhodne, jak pokračovat dále. Musí být rozhodnuto, zda ukončit tento projekt a přesunout se do fáze Implementace nebo zda zahájit další opakování některých fází nebo dokonce začít zcela nový data miningový projekt. Tento úkol zahrnuje analýzy zbývajících zdrojů a rozpočtu, které ovlivní rozhodnutí. Měly by být sepsány možné kroky společně s důvody pro a proti pro každou volbu a na závěr i rozhodnutí.

5.6 Využití výsledků

Pro nasazení data miningových modelů do obchodních činností bere tento úkol výsledky hodnocení a vyvozuje z nich strategii pro implementaci. Pokud byl identifikován obecný postup pro vytvoření platného modelu (modelů), je zde tento postup dokumentován pro pozdější použití.

Během zavádění modelů by nemělo být opomenuto vytvoření plánů pro kontrolu a údržbu. Jejich význam roste, pokud se výsledky data miningových analýz mají stát součástí každodenních aktivit organizace. Důkladná příprava strategie údržby pomáhá vyhnout se zbytečně dlouhým obdobím, po která jsou data miningové výsledky špatně užívány. Z důvodu kontroly nasazení výsledků je nutný detailní plán kontrolní činnosti. Tento plán bere do úvahy specifický typ nasazení.

Na konci projektu by měla být sepsána závěrečná zpráva. Ta může mít podobu buďto stručného shrnutí anebo může jít o závěrečné a vyčerpávající prezentování všech výsledků, jichž bylo dosaženo během celého procesu.

Závěrečným vypracovávaným dokumentem, který CRISP uvádí, je revize projektu. Někomu se možná může zdát, že tato zpráva je zbytečná, neboť se nepředává zákazníkovi ale slouží pro vnitřní potřeby data miningové firmy. Tato část by rozhodně neměla chybět. Jde o shromažďování podnikových znalostí (tzv. řízení znalostí). Pracovníci mají zhodnotit, co šlo dobře a co špatně, co bylo uděláno dobře a co je potřeba zlepšit. Shrnují důležité zkušenosti získané během projektu. Upozorňují na nebezpečná místa v analýze, na zavádějící přístupy nebo ukazatele pro výběr nejvhodnějších data miningových technik. Právě tyto zaznamenané individuální postřehy umožňují pracovníkům sdílet své zkušenosti a pracovat tak v dalších projektech efektivněji. [3]

6. Data mining v praxi

V praxi se data mining používá v mnoha oblastech, jako příklad je uveden data mining s CRM, neboli řízení vztahů se zákazníky.

6.1 Data mining a CRM

Oblast CRM představuje dnes nejčastější praktické případy pro data mining. Uveďme nejcharakterističtější příklady:

- Analýza prodeje (analýza nákupního chování, vyhodnocení speciálních nabídek, porovnání prodeje podle většího množství atributů, vyhodnocení sezónnosti atd.). Často citovaným příkladem z praxe je zjištění jednoho hypermarketu, že ke konci pracovního týdne se společně často prodává pivo a dětské pleny. Reorganizace prodejní plochy přinesla zvýšení prodeje těchto produktů s návratností nákladů vložených do analýzy (byla zde použita analýza nákupního košíku).
- Cílený marketing (nalezení nejlevnější cesty, jejíž prostřednictvím na určitou nabídku reaguje dostatečné množství respondentů). Praktická účinnost plošného marketingu bývá udávána max. v jednotkách procent. Po oslovení vzorku vybraného pomocí data mining technik lze výrazně ušetřit jednotkové náklady na marketing za minimálního snížení úspěšnosti celkové odezvy. Příkladem může být marketingová kampaň mobilního operátora, u které se zvýšila odezva z 2-3% na 15%.
- Analýza rizik, detekce zneužití (analýza finančních transakcí, analýza žádostí o půjčku, analýza pojistných událostí atd.). Společnosti vydávající kreditní karty již rutinně vytvářejí své modely pro autorizační procesy, obvykle s využitím neuronových sítí. Jiným příkladem může být Caterpillar, který automatizuje vyhodnocování žádostí o proplacení záručních oprav od smluvních servisů a podrobněji zkoumá jen ty, které se jeví jako neadekvátní.
- Vyhodnocení loajality, minimalizace pravděpodobnosti ztráty zákazníka (Kdo jsou naši nejlepší zákazníci? Jak se můžeme ubránit tomu, že přejdou ke konkurenci? Komu je vhodné nabídnout speciální podmínky nebo zákaznický program? Jaká je hodnota zákazníka v průběhu celého obchodního kontaktu s ním (LTV)?). Příkladem z praxe může být zjištění jedné banky, že lidé, kteří nevyužívají kreditu v průběhu roku, ale činí tak v listopadu a prosinci (k zajištění dodatečných vánočních výdajů) jsou vysoce ziskoví.

- Vývoj nového produktu, cross-selling. Příkladem může být banka, která s využitím shlukové analýzy identifikovala dříve nepovšimnutou skupinu, kde 39 procent zákazníků mělo jak osobní, tak podnikatelský účet a 11% procent z této skupiny žádalo o půjčku zajištěnou vlastní nemovitostí. Po zjištění, že vysoké procento těchto žádostí bylo ohodnoceno jako nízkorizikové banka začala pracovat s hypotézou, že nemálo jejich zákazníků zahajuje podnikání tak, že za úvodní půjčku ručí vlastní nemovitostí. Hypotéza byla ověřena dobrou odezvou po zavedení nové nabídky zaměřené na tuto skupinu. [11]

6.2 Praktické rozdíly mezi vyhledáváním v databázích a data miningem

6.2.1 Data-mining a modelování

Data-mining je technologie založena na automatickém vyhledávání dosud neobjevených souvislostí v datech. V oblasti CRM se jedná například o hledání trendů a vzorů týkajících se chování zákazníků.

Analytické nástroje ve spojení s data-miningem umožňují modelovat budoucí chování zákazníků. Vznikají modely znázorňující závislé proměnné, které představují pravděpodobnost určitého chování zákazníka. Závislé proměnné závisí na dalších závislých či nezávislých proměnných o zákazníkovi (demografická data, chování, reakce, využívané produkty). V datovém skladu se tyto modely také ukládají. Modely se podle potřeb přepočítávají, například pokud si zákazník objedná nový produkt či uzavře smlouvu.

Pojďme si teď na příkladu ukázat roli závislých a nezávislých proměnných na jednom praktickém modelu. Závislou proměnnou může být například "uzavření půjčky". S ní souvisí nezávislé proměnné, jako je věk, pohlaví, vzdělání, výše příjmu, lokalita a další. Na základě těchto informací můžeme pomocí analýz zjistit, které osoby určitého věku, pohlaví, příjmu či lokality mohou být pro firmu těmi nejlepšími zákazníky. Ovšem zjištění závislostí mezi daty je poměrně obtížné, protože musíme respektovat i další souvislosti týkající se klienta. Například pro banku jsou klíčové charakteristiky týkající se pohybu finančních prostředků na účtu, četnost výběru z bankomatů a jejich výše, využívání platebních karet při placení v obchodech atd.

Analýzy jsou vhodné také pro zjišťování toho, jaké produkty si zákazníci kupují společně. Pokud se odhalí výraznější výskyt tohoto jevu, můžeme pak dalším zákazníkům nabízet

rovnou i jiné zboží, než je pouze právě to, které si kupují. Zvyšuje se zde pravděpodobnost toho, že si zákazník koupí i další zboží, jelikož mu bylo nabídnuto "na míru".

Existují také způsoby analýzy, které odhalí, jak nejlépe nabídnout daný produkt určitému zákazníkovi (resp. skupině zákazníků s podobnými charakteristikami). Reaguje zákazník na slevy? Motivuje ho ke koupi možnost získání dárku? Je zákazník zaměřen spíše na produkty spadající do vyšší či nižší cenové kategorie? Na tyto a další podobné otázky by mělo být kvalitně sestavené analytické CRM schopno odpovědět.

Důležitá je také diferenciace strategie přístupu k zákazníkům. Ne každý je pro společnost stejně ziskový. S tím souvisí potřeba minimalizovat náklady na málo ziskové či neziskové zákazníky. V této souvislosti používaný termín "Celoživotní hodnota zákazníka" (Customer Lifetime Value) zahrnuje současný i budoucí přínos pro danou firmu. Zjišťuje se pomocí data-miningu, predikcí a modelů vzájemných vztahů. Snahou je odhalit zákazníky, kteří přinášejí nejvyšší zisky a relativně malé riziko oproti jiným. O tyto zákazníky (i potenciální) by se měly firmy nejvíce zajímat, získávat je a bránit jim v tom, aby přecházeli ke konkurenci. Opět je zde platné Paretovo pravidlo 20:80, které nám říká, že 80 procent zisku nám přináší pouze 20 procent zákazníků. Právě na ty je třeba se zaměřit a konkurenci přenechat ty ostatní. Hodnotou zákazníka se zabývá Customer Value Management (CVM), tedy koncept řízení hodnoty zákazníka. [12]

Dokonce jsem se setkal s názorem, že se data mining bude využívat v souvislosti s odhalením teroristického útoku.

6.2.2 Vyhledávání v databázích

Oblast použití transakčních databázových systémů je skoro neomezená. Primárním cílem při jejich návrhu je umožnit klientům databázového serveru vykonávání velkého množství transakcí online, například bankovních, obchodních a podobně. Cílem transakčních databázových systému je automatizace každodenních činností, které jsou předmětem našeho podnikání, například skladové hospodářství, mzdy, nákup a prodej, případně řízení a monitorování technologických procesů v reálném čase. [10]

Dnešní společnosti sestavují své informace do databází – databází zákazníků, databází výrobků, databází obchodních zástupců a pak slučují údaje z různých databází. Databáze zákazníků obsahují například jméno každého zákazníka, jeho adresu, minulé transakce a

v některých případech rovněž jeho demografické a psychologické rysy (aktivity, zájmy, názory). Místo aby společnost podnikala „kobercový nálet" a rozeslala novou nabídku všem zákazníkům ve své databázi, osloví jen určité zákazníky podle množství jejich nákupů v poslední době, frekvence nákupů a jejich finanční hodnotě. Pošle nabídku jen těm zákazníkům, kteří dosáhnou nejlepších výsledků. Kromě toho, že ušetří na poštovném, často dosáhne dvouciferné procentuální míry reakce.

Například Pizza Hutt tvrdí, že má ze všech společností poskytujících rychlé občerstvení největší databázi zákazníků, v níž je 40 miliónů domácností v USA – neboli 40 – 50% trhu. Z jednotlivých provozoven se shromažďují zprávy o zákaznících. Pizza Hutt může třídit informace podle oblíbených náplní pizzy jednotlivých zákazníků, data poslední objednávky nebo podle toho, zda si objednáváte k pizze s feferonkami salát. S pomocí Teradata Warehouse Miner dokáže Pizza Hutt nejen ze svých directmailových kampaní odstranit nákladné duplikáty, ale zaměřit svůj marketing tak, aby nalezl nejlepší kuponové nabídky pro každou domácnost a dokázal předpovídat úspěch svých kampaní.

Společnosti tyto údaje ukládají a umožňují lidem, kteří činí rozhodnutí, získat k nim snadný přístup. Navíc najímáním analytiků znalých sofistikovaných statistických metod dokážou z údajů „těžit" a získávat nové nápady vztahující se k zanedbávaným zákaznickým segmentům, posledním zákaznickým trendům a podobně. Informace o zákaznících lze různými způsoby spojovat s informacemi o výrobcích a informace od obchodních zástupců a vytěžit z těchto spojení další nápady. K účinnému a vhodnému řízení všech rozdílných databází stále více firem nyní používá business intelligence software.

Společnost Wells Fargo například používáním vlastní technologie získala schopnost sledovat a analyzovat veškeré bankovní transakce svých 10 milionů maloobchodních zákazníků – ať již z bankomatů, bankovních poboček nebo internetového bankovnictví. Spojí– li si společnost Wells Fargo transakční data s osobními daty poskytnutými zákazníky, může přijít s cílenými nabídkami, které se shodují s nějakou zásadní změnou v životě určitého zákazníka. V důsledku toho prodá Wells Fargo čtyři bankovní produkty na zákazníka, zatímco průměr v jejím odvětví činí 2,2 produktu na zákazníka. [13]

7. Závěr

Obsahem této bakalářské práce bylo na základě teoretických poznatků uvést a porovnat jednotlivé způsoby a možnosti využití data miningu. Jednalo se o mou první zkušenost s data miningem a myslím že tato práce mi dala nové zkušenosti a poznatky v této oblasti.

Členění data miningu je velmi rozsáhlé tzn., že existuje mnoho metod, jak dolovat data, proto jsem vybral do této bakalářské práce pouze několik metod dolování dat. V první řadě to jsou neuronové sítě. Jsou to systémy, které provádějí rozpoznání vzorů v přijatých vstupech na základě modelů, jak zpracovávají informace neurony savců. Dále to jsou rozhodovací stromy, které jsou analytické nástroje sloužící k nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně. Nebo logistická regrese a ta je velmi užitečná v marketingu, protože se často snažíme předvídat diskrétní akci, například odezvu na nabídku nebo nesplacení půjčky.

Jedním z velmi významných možností využití data miningu pro podnik je rozeslání nabídky na produkt pouze těm odběratelům, kteří odebrali nějaký produkt v poslední době, jsou to pro nás klíčoví odběratelé a je možnost jim poskytnout nějakou slevu. Nevyplatí se rozesílání nabídky všem zákazníkům a to především z finančního hlediska. V tomto se využívá paretovo pravidlo 20:80, které říká, že 80 % zisku přináší pouze 20 % zákazníků, proto je nutno si těchto zákazníků vážit.

Data mining se z mého úhlu pohledu nejvíce rozvíjí v oblasti obchodování, ve finanční a marketingové sféře. Obchodníci jej využívají k předpovídání spotřebitelských vzorců chování zákazníků a společnosti podnikající v oblasti kreditních karet k odhalování podvodů. Naopak nejpomaleji se patrně data mining rozmáhá v exaktních vědách, a to díky nižší transparentnosti speciálních metod data mining a obtížné porovnatelnosti výsledků různých programů data miningu.

Rychlý rozvoj také zaznamenává v průmyslu (modelování spolehlivosti), nebo také v lékařství. Obecně bych snad mohl říci, že všude tam, kde na hledaný výsledek působí celá řada faktorů a kde často potřebujeme nalézt více řešení (ať už pomocí různých metod nebo při různých počátečních podmínkách), které chceme navzájem srovnávat.

Data mining je poměrně mladá metoda, která se praxi začala využívat teprve na přelomu 70. a 80. let 20. století. Myslím, že data mining není ještě ve svém vývoji úplně na konci a že se bude tato technologie ještě zdokonalovat.

Použitá literatura

[1] TVRDÍKOVÁ M. *Aplikace moderních informačních technologií v řízení firmy* 1. vyd. Praha 2008.

[2] CHROMÝ J. *Elektronické podnikání*, 2 přepracované vydání. Praha 8, 2009.

[3] URL: < http://datamining.xf.cz/view.php?cisloclanku=2002102809> [cit. 2002-10-28]

[4] HAN J. – KAMBER M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001, ISBN 1-55860-489-8.

[5] ENDORF C., SCHULTZ E., MELLANDER J. : Hacking – *detekce a prevence počítačového útoku*, 1. vyd. Grada 2007. 356 s., ISBN 80-247-1035-8

[6] RUD, Olivia. *Data mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. 1. vydání. Praha: Computer Press, 2001. 370 s. ISBN 80-7226-577-6.

[7] KOHONEN T, *Self-organization of very large document collections: State of the art*. Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, volume 1, pages 65-74. Springer, London, 1998

[8] URL: < http://euromise.vse.cz/kdd/index.php?page=proceskdd> [cit. 2002-10-24]

[9] DUNNIGAN F., James. *Bojiště zítřka: tváří v tvář globální hrozbě kybernetickému útoku*. 1. vydání. Praha: BARONET, 2004. 359 s. ISBN 80-7214-642-4.

[10] LACKO Luboslav. *Datové sklady analýza OLAP a dolování dat.* Brno: Computer Press, a.s., 2003. ISBN 80-7226-969-0

[11] BERRY Michael and LINOFF Gordon. *Dat Mining Techniques: For Marketing*, Sales and Customer Support. New York: Wiley Computer Publishing, 1997.

[12] URL: < http://deathless.blog.cz/0805/analyticke-crm>

[13] KOTLER, Philips, KELLER, Kevin Lane, *Marketing management*, 12. vydání, Praha: Grada Publishing, 2007, ISBN 978-80-247-1359-5

Seznam obrázků

obr. 1 Řetězec dodávky dat

obr. 2 Jednoduché schéma business intelligence

obr. 3 Klasifikační strom odpovědí

obr. 4 Logistická regrese

obr. 5 Shluky dokumentů jako Kohonenova mapa

obr. 6 Metodologie CRISP-DM